

Ab-Initio Interview Questions For Beginners

07/01/2011

<AMEER BASHA SHAIK>

ameerbasha.shaik@tcs.com

Why we go for Ab-Initio?

Answers:

- Ab-Initio designed to support largest and most complex business applications.
- We can develop applications easily using GDE for Business requirements.
- Data Processing is very fast and efficient when compared to other ETL tools.
- Available in both Windows NT and UNIX.

Differences Between Ab-Initio and Informatica?

Answers:

- Informatica and Ab-Initio both support parallelism. But Informatica supports only one type of parallelism but the Ab-Initio supports three types of parallelisms.
 - Component
 - Data Parallelism
 - Pipe Line parallelism.
- We don't have scheduler in Ab-Initio like Informatica , you need to schedule through script or you need to run manually.
- Ab-Initio supports different types of text files means you can read same file with different structures that is not possible in Informatica, and also Ab-Initio is more user friendly than Informatica .
- Informatica is an engine based ETL tool, the power this tool is in it's transformation engine and the code that it generates after development cannot be seen or modified.
- Ab-Initio is a code based ETL tool, it generates ksh or bat etc. code, which can be modified to achieve the goals, if any that can not be taken care through the ETL tool itself.
- Initial ramp up time with Ab-Initio is quick compare to Informatica, when it comes to standardization and tuning probably both fall into same bucket.
- Ab-Initio doesn't need a dedicated administrator, UNIX or NT admin will suffice, where as Informatica need a dedicated administrator.
- With Ab-Initio you can read data with multiple delimiter in a given record, where as Informatica force you to have all the fields be delimited by one standard delimiter.

Ab-Initio Interview Questions For Beginners

- Error Handling - In Ab-Initio you can attach error and reject files to each transformation and capture and analyze the message and data separately. Informatica has one huge log! Very inefficient when working on a large process, with numerous points of failure.

What are the most commonly used components in a Ab-Initio graphs?

Answers:

- input file / output file
- input table / output table
- lookup / lookup_local
- reformat
- gather / concatenate
- join
- run sql
- join with db
- compression components
- filter by expression
- sort (single or multiple keys)
- rollup
- partition by expression / partition by key

How do we handle if DML changing dynamically?

Answers:

There are lot many ways to handle the DMLs which changes dynamically with in a single file.

Some of the suitable methods are to use a **conditional DML** or to **call the vector functionality** while calling the DMLs.

What is meant by limit and ramp in Ab-Initio? Which situation it's using?

Answers:

The limit and ramp are the variables that are used to set the reject tolerance for a particular graph. This is one of the option for reject-threshold properties. The limit and ramp values should pass if enables this option.

Graph stops the execution when the number of rejected records exceeds the following formula..

$\text{limit} + (\text{ramp} * \text{no_of_records_processed})$.

The default value will be set to 0.0.

The limit parameter contains an integer that represents a number of reject events The ramp parameter contains a real number that represents a rate of reject events in the number of records processed.

Typical Limit and Ramp settings

Limit = 0 Ramp = 0.0 Abort on any error

Limit = 50 Ramp = 0.0 Abort after 50 errors

Limit = 1 Ramp = 0.01 Abort if more than 2 in 100 records causes error

Limit = 1 Ramp = 1 Never Abort

What is mean by Layout?

Answers:

A layout is a list of host and directory locations, usually given by the [URL](#) of a [file](#) or [multi file](#). If a layout has multiple locations but is not a multi file, the layout is a list of URLs called a [custom layout](#).

A program component's layout is the list of hosts and directories in which the component runs.

A dataset component's layout is the list of hosts and directories in which the data resides. Layouts are set on the [Properties Layout tab](#).

The layout defines the level of [Parallelism](#) . Parallelism is achieved by partitioning data and computation across processors.

What are Cartesian joins?

Answers:

A Cartesian join will get you a Cartesian product. A Cartesian join is when you join every row of one table to every row of another table. You can also get one by joining every row of a table to every row of itself.

What are the uses of is_valid, is_define functions?

Answers:

is_valid and **is_defined** are Pre defined functions

is valid(): Tests whether a value is valid.

The **is_valid** function returns:

- The value **1** if expr is a valid data item.
- The value **0** if the expression does not evaluate to NULL.

If expr is a record type that has field-validity checking functions, the **is_valid** function calls each field validity checking function. The **is_valid** function returns **0** if any field-validity checking function returns **0** or NULL.

Example:

```
is_valid(1) 1
```

```
is_valid("oao") 1
is_valid((decimal(8))"1,000") 0
is_valid((date("YYYYMMDD"))"19960504") 1
is_valid((date("YYYYMMDD"))"abcdefgh") 0
is_valid((date("YYYY-MMM-DD"))"1996-May-04") 1
is_valid((date("YYYY-MMM-DD"))"1996*May&04") 0
```

is_defined():

Tests whether an expression is not NULL.

The is_defined function returns:

The value 1 if expr evaluates to a non-NULL value.

The value 0 otherwise.

The inverse of is_defined is is_null.

What is meant by merge-join and hash-join? Where those are used in Ab-Initio?

Answers:

The command-line syntax for Join Component consists of two commands. The first one calls the component, and is one of two commands:

- mp merge-join to process sorted input
- mp hash-join to process unsorted input

How does force_error function work ? If we set never abort in reformat , will force_error stop the graph or will it continue to process the next set of records ?

Answers:

force_error as the name suggests it works on as to force an error in case of not meeting of any conditions mentioned.

The function can be used as per the requirement.

If you want to stop execution of graph in case of not meeting a specific condition say you have to compare the input and out put records reconciliation and the graph should fail if the input record count is not same as output record count

"THEN set the reject-threshold to Abort on first reject" so that the graph stops.

Note:- force_error directs all the records meeting the condition to reject port with the error message to error port.

In certain special circumstances you can also use to treat the reject port as an additional data flow path leaving the component.

When using force_error to direct valid records to the reject port for separate processing you must remember that invalid records will also be sent there.

When using force_error for this purpose "set the reject-threshold to Never Abort" so that the graph does not fails and meets the purpose.

What is the use of unused port in join component?

Answers:

While joining two input flows, records which match the join condition goes to output port and we can get the records which do not meet the join condition at unused ports.

What is meant by dedup Sort with null key?

Answers:

If we don't use any key in the sort component while using the dedup sort, then the output depends on the **keep** parameter. It considers whole records as one group.

- first - only the first record
- last - only last record
- unique_only - there will be no records in the output file.

How many numbers of inputs join component support ?

Answers:

Join will support maximum of 60 inputs and minimum is 2 inputs.

What is max-core? What are the Components that use MAX_CORE?

Answers:

The value of the MAX_CORE parameter is that it determines the maximum amount of memory, in bytes, that a specified component will use. If the component is running in parallel, the value of MAX_CORE represents the maximum memory usage per partition. If MAX_CORE is set too low the component will run slower than expected. Too high and the component will use too many machine resources and slow up Dramatically.

The Max core parameter can be defined in the following components:

- SCAN
- in-memory SCAN
- ROLLUP
- in-memory ROLLUP
- in-memory JOIN
- SORT

Whenever these components are used and have the component set to parameter set to "In-memory; Inputs need not be sorted", a max-core variable must be specified.

How many types of joins are in Ab-Initio?

Answers:

Description of Join Types:

Join is based on a match key for inputs, Join components describes out port, unused ports, reject ports and log port.

Inner Joins

The most common case is when join-type is Inner Join. In this case, if each input port contains a record with the same value for the key fields, the transform function is called and an output record is produced.

If some of the input flows have more than one record with that key value, the transform function is called multiple times, once for each possible combination of records, taken one from each input port.

Whenever a particular key value does not have a matching record on every input port and Inner Join is specified, the transform function is not called and all incoming records with that key value are sent to the unused ports.

Full Outer Joins:

Another common case is when join-type is Full Outer Join: if each input port has a record with a matching key value, Join does the same thing as it does for an Inner Join.

If some input ports do not have records with matching key values, Join applies the transform function anyway, with NULL substituted for the missing records. The missing records are in effect ignored.

With an Outer Join, the transform function typically requires additional rules (as compared to an Inner Join) to handle the possibility of NULL inputs.

Explicit Joins:

The final case is when join-type is Explicit. This setting allows you to specify True or False for the record-required n parameter for each in n port. The settings you choose determine when Join calls the transform function.

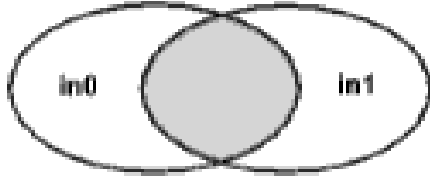
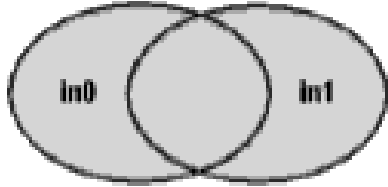
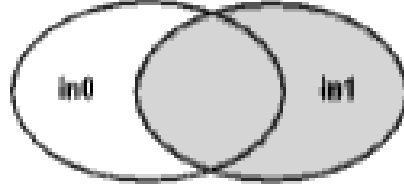
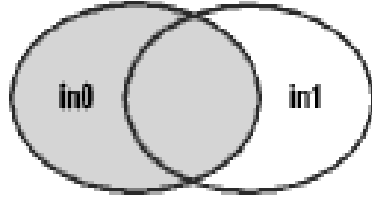
The join-type and record-required n Parameters

The two intersecting ovals in the diagrams below represent the key values in the records on the two ports — **in0** and **in1** — that are the inputs to join:

For each possible setting of **join-type** or (if **join-type** is **Explicit**) combination of settings for

record-required n, the shaded region of each of the following diagrams represents the inputs for which Join calls the transform. Join ignores the records that have key values represented by the white regions, and consequently those records go to the **unused** port.

Ab-Initio Interview Questions For Beginners

Case 1 Inner Join join-type	
Case 2 Full Outer Join join-type	
Case 3a (Left Outer join) Explicit join-type: Record-required0: False Record-required1: True	
Case 3b (Right Outer join) Explicit join-type: Record-required0: True Record-required1: False	

What is semi-join?

Answers:

Ab-Initio gives 3 examples of joins: inner join, outer join, and semi join.

- for inner join '**record_required N**' parameter is true for all "in" ports.
- for outer join it is false for all the "in" ports.
- for semi join it is true for both port (like Inner Join), but the de-dup option is set only on one side

When we use Dynamical DML?

Answers:

Dynamic DML is used if the input meta data can change. Example: at different time different input files are received for processing which have different dml. in that case we can use flag in the dml and the flag is first read in the input file received and according to the flag its corresponding dml is used.

Explain the differences between Replicate and BROADCAST?

Answers:

Replicate takes records from input flow arbitrarily combines and gives to components which connected to its output port.

Broadcast is partition component copies the input record to components which connected to its output port.

Consider one example, input file contains 4 records and level of parallelism is 3 then Replicate gives 4 records to each component connected to its out port whereas Broadcast gives 12 records to each component connected to its out port.

Explain the difference between REFORMAT and Redefine FORMAT?

Answer:

Reformat changes the record format by adding or deleting fields in the DML record.

Length of the record can be changed.

Redefine copies its input flow to its out port without any transform.

Redefine is used to rename the fields in the DML. But Length of record should not change.

How to work with parameterized graphs?

Answers:

Parameterized graphs specifies everything through parameters. i.e, data locations in input/output files, DMLs etc...

What is driving port? When do you use it?

Answers:

When joining inputs (in0, in1, ...) one of the ports is used as "driving (by default - in0). Driving input is usually the largest one. Whereas the smallest can have "Sorted-Input" parameter be set to "Input need not be sorted" because it will be loaded completely in memory. This concept comes to existence only when we use Sorted-Input parameter as [In-Memory join](#).

How can we test the ab-Initio manually and automation?

Answers:

By running a graph through GDE is manual test.

By running a graph using deployed script is automated test.

What is error called 'depth not equal'?

Answers:

When two components are linked together if their layout does not match then this problem can occur during the compilation of the graph. A solution to this problem would be to use a partitioning component in between if there was change in layout.

What is the function you would use to transfer a string into a decimal?

Answers:

For converting a string to a decimal we need to typecast it using the following syntax,

```
out.decimal_field :: ( decimal( size_of_decimal ) ) string_field;
```

The above statement converts the string to decimal and populates it to the decimal field in output.

Which one is faster for processing fixed length dmls or delimited dmls and why?

Answers:

Fixed length,because for delimited dml it has to check for delimiter every time but for fixed length dml directly length will be taken.

What are kinds of layouts does ab-Initio supports?

Answers:

Ab-Initio supports two kinds of Layouts:

- Serial Layout
- Multi layout.

In Ab-Initio Layout tells which component should run where and it also gives level of parallelism.

For serial Layout,level of parallelism is 1.

For Multi layout,Level of parallelism depends on data partition.

How can you run a graph infinitely?

Answers:

To run a graph infinitely,

- The end script of the graph should call the .ksh file of the graph. Thus if the name of the graph is abc.mp then in the end script of the graph there should be a call to abc.ksh. Then this graph will run infinitely.
- Run the deployed script in a loop infinitely.

How can I calculate the total memory requirement of a graph?

Answers:

- You can roughly calculate memory requirement as:
- Each partition of a component uses:~ 8 MB + max-core (if any)
- Add size of lookup files used in phase (if multiple components use same lookup only count it once)
- Multiply by degree of parallelism. Add up all components in a phase; that is how much memory is used in that phase.
- Add size of input and output datasets(Total memory requirement of a graph) > (the largest-memory phase in the graph).

What is multistage component?

Answers:

Multistage component are nothing but the transform components where the records are transformed into five stages like input selection, temporary records initialization, processing , finalization and output selection.

examples of multistage components are like

- Rollup
- Scan
- Normalize
- Denormalize sorted.

what is the use of aggregation when we have rollup as we know rollup component in ab-Initio is used to summarize group of data record. then where we will use aggregation ?

Answers:

Rollup has a good control over record selection grouping and aggregation as compared to that of aggregate. Rollup is an updated version of aggregate.

When Rollup is in template mode ,it has aggregation functions to use. So it is better to go for Rollup.

Phase verses Checkpoint ?

Answer:

Difference between a phase and checkpoint .

phases are used to break up a graph so that it does not use up all the memory , it limits the number of active components thus reduce the number of components running in parallel hence improves the performance .

Phases make possible the effective utilization of the resources such as memory disk space and CPU So when we have memory consuming components in the straight flow and the data in flow is in millions we can separate the

process out in one phase so as the CPU allocation is more for the process to consume less time for the whole process to get over.

Temporary files created during a phase will be deleted after completion of that phase.

Don't put phase after Replicate,sort,across all to all flows and temporary files.

Check points are used for the purpose of recovery.

In contrary Checkpoints are like save points .These are required if we need to run the graph from the saved last phase recovery file(phase break checkpoint) if it fails unexpectedly.

At job start,output datasets are copied into temporary files and after the completion of check pointing all datasets and job state are copied into temporary files. so if any failure occurs job can be run from last committed check point.

Use of phase breaks which includes the checkpoints would degrade the performance but ensures save point run.

The major difference between these two is that phasing deletes the intermediate files made at the end of each phase as soon as it enters the next phase.

On the other hand what check pointing does is...it stores these intermediate files till the end of the graph. Thus we can easily use the intermediate file to restart the process from where it failed. But this cannot be done in case of phasing.

We can have phases without check points.

We can not assign checkpoints without phases.

In Ab-Initio, How can you display records between 50-75.. ?

Answers:

Input dataset having 100 records. I want records between 50-75 then use

```
m_dump <dml> <mfs file> -start 50 -end 75
```

For serial and mfs there are many ways the components can be used.

- 1.Filter by Expression : use next_in_sequence() >50 && next_in_sequence() <75
2. We can also use multiple LEADING RECORDS components for meeting the requirement.

If you have the access to Co>Op then you can try an alternate.

Say suppose the input file is : file 1

Use the Run program component in GDE and write the below command:

```
`sed -n50 75p file 1 > file 2`
```

What is the order of evaluation of parameters?

Answers:

When you run a graph, parameters are evaluated in the following order:

- The host setup script is run.

Ab-Initio Interview Questions For Beginners

- Common (i.e, included) sandbox parameters are evaluated.
- Sandbox parameters are evaluated.
- The **project-start.ksh** script is run.
- Graph parameters are evaluated.
- The graph Start Script is run.
- The execution of process is run simultaneously based component's layouts.
- The Lookup files is run
- The graph Meta data is checking process.
- The in/out file paths with files are checking.
- The graph runs as order of phase0, phase1, phase2, ...

How do you convert 4-way MFS to 8-way mfs?

Answers:

By partitioning, we can use any partition method to partition.

Partitioning methods are:

- Partition by Round-robin
- Broadcast
- Partition by Key
- Partition by Expression
- Partition by Range
- Partition by Percentage
- Partition by Load Balance

For data parallelism,we can use partition components. For component parallelism,we can use replicate component.

Like this which component(s) can we use for pipeline parallelism?

Answers:

When connected sequence of components of the same branch of graph execute concurrently is called pipeline parallelism.

Components like reformat where we distribute input flow to multiple o/p flow using output index

depending on some selection criteria and process those o/p flows simultaneously creates pipeline parallelism.

But components like sort where entire i/p must be read before a single record is written to o/p can not achieve pipeline parallelism.

What is multi directory?

Answers:

A multi directory is a parallel directory that is composed of individual directories, typically on different disks or computers. The individual directories are partitions of the multi directory. Each multi directory contains one control directory and one or more data directories. Multi files are stored in multi directories.

What is multi file?

Answers:

A multi file is a parallel file that is composed of individual files, typically on different disks or computers. The individual files are partitions of the multi file. Each multi file contains one control partitions and one or more data partitions. Multi files are stored in distributed directories called multi directories. This diagram shows a multi directory and a multi file in a multi file system:

The data in a multi file is usually divided across partitions by one of these methods:

- Random or round robin partitioning
- Partitioning based on ranges or functions
- Replication or broadcast, in which each partition is an identical copy of the serial data.

What is mean by GDE, SDE? What is purpose of GDE, SDE?

Answers:

- GDE - Graphical Development Environment –it is used for developing the graphs
- SDE – Shell Development Environment, which is used for developing the korn shell script on co>operating system.

What is difference between Rollup and Scan ?

Answers:

Roll up comp:

Rollup evaluates a group of input records that have the same key and then generates data records that either summarize each group or select certain information from each group.

Using Rollup component can evaluates to two ways as follows: 1. Template mode 2. Expanded Mode

1. Template Mode:

This mode options evaluates using built aggregation functions alike sum, min, max, count, avg, product, first, last.

2. Expanded Mode:

This mode option can evaluates using (without aggregation functions) user defined functions alike temporary function, initialize, finalize and rollup functions in transform function propriety.

Scan comp:

Scan generates a series of cumulative summary records — such as successive year-to-date totals for groups of data records. Scan produces intermediate summary records.

Rollup is for group by and Scan is for successive total. Basically, when we need to produce summary then we use scan. Rollup is used to aggregate data.

What is Runtime Behavior of Rollup?

Answers:

Roll up can supports two types of modes.

1.Template Mode:

This mode options evaluates using built aggregation functions alike sum, min, max, count, avg, product, first, last.

2. Expanded Mode:

This mode option can evaluates using (without aggregation functions) user defined functions alike temporary function, initialize, finalize and rollup functions in transform function propriety.

Rollup component's performance differs from using **Rollup Input is Sorted** and **Rollup Input is Unsorted**

When Rollup Input is sorted

When you set the **sorted-input** parameter to **Input must be sorted or grouped** (the default), Rollup requires data records grouped according to the **key** parameter. If you need to group the records, use Sort with the same key specifier that you use for Rollup. It will produces sorted outputs in output port.

When Rollup Input is Unsorted

When you set the **sorted-input** parameter to **In memory: Input need not be sorted**, Rollup accepts un grouped input, and groups all records according to the **key** parameter. It does not produce sorted output.

How do you do rollback in Ab-Initio?

Answers:

Ab-Initio has supports very good recovery options for any failures at runtime and interrupted powers at development time.

Development time:

You can get a recovery graph file while occurred any interrupted failures at development time.

At Runtime:

You can get a recovery file while occurred any failures at execution of graph and you can restart the execution. The recovery file has last checkpoint information and restarts from last checkpoint onwards.

you can use two ways to rollback the Ab-Initio graphs

m_rollback -d -deletes all intermediate files and checkpoints

What is internal execution (process) of the Ab-Initio graphs in Ab-Initio co>operating system on while running the graphs?

Answers:

Normally the Ab-Initio Co> operating system checks relevant code compatible of GDE and

Co>operating system. if you are used any lookup files in graphs. This is called lookup layout checking.

The graphs are having input and output files and it checks whether the path are correct or not, given below the sequence of process has done while running the graphs.

- Checks lookup files layouts.
- Checks meta data part (this is part check whether data types are used or not and related

everything) – dml checking for each component basis.

- Checks input files
- Checks output files
- Checks each component's layouts
- Finally, it checks flow of process assigns to straight.

What does dependency analysis mean in Ab-Initio?

Answers:

dependency analysis will answer the questions regarding data lineage that is where does the data comes from and what applications produced depend on this data etc..

What is meant by Fencing in Ab-Initio?

Answers:

In Software World fencing means job controlling on priority basis.

In Ab-Initio it actually refers to customized phase breaking.

A well fenced graph means no matter what is source data volume process will not cough in dead locks.

It actually limits the number of simultaneous processes.

In Ab-Initio you need to Fence the job in some times to stop the schedule.

Fencing is nothing but changing the priority of the particular job.

What is the function of fuse component?

Answers:

Fuse combines multiple input flows into a single output flow by applying a transform function to corresponding records of each flow

Runtime Behavior of Fuse

Fuse applies a transform function to corresponding records of each input flow. The first time the transform function executes, it uses the first record of each flow. The second time the transform function executes, it uses the second record of each flow, and so on. Fuse sends the result of the transform function to the out port.

The component works as follows. The component tries to read from each of its input flows.

- If all of its input flows are finished, Fuse exits.
- Otherwise, Fuse reads one record from each still-unfinished input port and a NULL from each finished input port.

what is data skew? how can you eliminate data skew while i am using partition by key?

Answers:

The skew of a data or flow partition is the amount by which its size deviates from the average partition size expressed as a percentage of the largest partition:

Skew of data $(\text{partition size} - \text{avg.partition size}) * 100 / (\text{size of largest partition})$

What is \$mpjret? Where it is used in ab-Initio?

Answers:

\$mpjret gives the status of a graph.

U can use \$mpjret in end script like

```
if 0 -eq($mpjret)
```

```
then
```

```
echo success
```

```
else
```

```
mailx -s [graph_name] failed mail_id
```

Difference between conventional loading and direct loading ? when it is used in real time ?

Answers:

Conventional Load:

Before loading the data all the Table constraints will be checked against the data.

Direct load:(Faster Loading)

All the Constraints will be disabled. Data will be loaded directly. Later the data will be checked against the table constraints and the bad data won't be indexed.

api conventional loading

utility direct loading.

How do you done the unit testing in Ab-Initio? How will you perform the Ab-Initio Graphs executions? How will you increase the performance in Ab-Initio graphs?

Answers:

The Ab-Initio Co>operating system is handling the graph with multiple processes running simultaneously. This is primary performance. Follows the given below actions:

1. The data separators mostly use “\307” and “\007” instead of “~”, “,” and special characters and avoids these delimiters. Because of the Ab-Initio has predefined these data separators.
2. Avoids repeated aggregation in graphs. You calculate for required aggregation at once and stores in file calls value using parameters and then you can use this parameter where it required.
3. Avoids the maximum number of components in graph and max core components in graphs.
4. Don't write any kinds looping statements in start script
5. Mostly use the sources are flat files

How do you improve the performance of a graph?

Answers:

There are many ways the performance of the graph can be improved.

- Use a limited number of components in a particular phase
- Use optimum value of max core values for sort and join components
- Minimize the number of sort components
- Minimize sorted join component and if possible replace them by in-memory join/hash join
- Use only required fields in the sort, reformat, join components
- Use phasing/flow buffers in case of merge, sorted joins

Ab-Initio Interview Questions For Beginners

- If the two inputs are huge then use sorted join, otherwise use hash join with proper driving port
- For large dataset don't use broadcast as partitioner
- Minimize the use of regular expression functions like re_index in the transfer functions
- Avoid repartitioning of data unnecessarily

How would you do performance tuning for already built graph?

Answers:

Steps to performance Tuning for already built graph.

- Understand the functionality of the Graph.
- Modularize(i.e,check for dependencies among components).
- Give Phasing.
- Check for correct Parallelism.
- Check for DB component(i.e,take required data from DB. Instead of taking whole data from DB which consumes more time and memory).

What is the difference between DB-config file and CFG file?

Answers:

A .dbc file has the information required for Ab-Initio to connect to the database to extract or load tables or views.

.dbc file contains:

- 1)Database name
 - 2)Database version
 - 3)Database nodes
- user name,password etc..

While .CFG file is the table configuration file created by db_config while using components like Load DB Table.

.cfg file contains:

- 1)Name of the Remote machine
- 2)User name/password to be used while connecting to DB
- 3)Location of the operating system on the remote machine.
- 4)The connection method.

It can be used to configure anything.

For Example,You want to set a value of a variable N,export its value and set the value in the .cfg file and run the .cfg file in start script.

What is .abinitiorc ? What it contain?

Answers:

.abinitiorc is a file which contains the credentials to connect to host.

Credentials like

1)Host IP

2)User-name

3>Password etc...

This is a config file for ab-Initio - in user's home directory and in \$AB_HOME/Config. It sets Ab-Initio home path, configuration variables (AB_WORK_DIR, AB_DATA_DIR, etc.), login info (id, encrypted password), login methods for hosts for execution (like EME host, etc.), etc.

Why might you create a stored procedure with the 'with recompile' option?

Answers:

Recompile is useful when the tables referenced by the stored procedure undergoes a lot of

modification/deletion/addition of data. Due to the heavy modification activity the execute plan

becomes outdated and hence the stored procedure performance goes down. If we create the stored procedure with recompile option, the sql server wont cache a plan for this stored procedure and it will be recompiled every time it is run.

What is the purpose of having stored procedures in a database?

Answers:

Main Purpose of Stored Procedure for reduce the network traffic and all sql statement executing in cursor so speed too high.

We use **Run SQL** and **Join with DB** components to run Stored Procedures.

What is mean by Co>Operating system and why it is special for Ab-Initio?

Answers:

Co > Operating System:Layered top to the Native operating system.

It converts the Ab-Initio specific code into the format, which the UNIX/Windows can understand and feeds it to the native operating system, which carries out the task.

How to retrieve data from database to source in that case which component is used for this?

Answers:

To unload (retrieve) Data from the database DB2, Informix, or Oracle we have components like **Input Table** and **Unload DB Table** by using these two components we can unload data from the database.

Input Table Component use the following parameters:

- 1)db_config file(which contains credentials to interface with Database)
- 2)Database Types
- 3)SQL file (which contains sql queries to unload data from table(s)).

How to execute the graph from start to end stages?Tell me and how to run graph in non Ab-Initio system?

Answers:

There are so many ways to do this,

- 1.you can run components according to phases how you defined.
- 2.by creating ksh, sh scripts also you can run.

What is Join With DB?

Answers:

Join with DB Component joins records from the flow or flows connected to its **in** port with records read directly from a database, and outputs new records containing data based on, transform function.

How do you truncate a table?

Answers:

Use Truncate Table component to truncate a table from DB in Ab-Initio.

Truncate Table Component has the following parameters:

- 1)db_config file(which contains credentials to interface with Database)
- 2)Database Types
- 3)SQL file (which contains sql queries to truncate table(s)).

Can we load multiple files?

Answers:

Yes,we can load multiple file in Ab-Initio.

What is the syntax of m_dump command?

Answers:

m_dump command prints the data in a formatted way.

The general syntax is

m_dump <dml> <file.dat>

"m_dump meta data data [action] "

e.g

m_dump emp.dml emp.dat -start 10 -end 20

– it will give record from 10 to 20 from emp.dat file.

How to Create Surrogate Key using Ab-Initio?

Answers:

A surrogate key is a substitution for the natural primary key.

–It is just a unique identifier or number for each record like ROWID of an Oracle table.

Surrogate keys can be created using

1)next_in_sequence

2)this_partition

3)no_of_partitions

Can any one give me an example of real-time start script in the graph?

Answers:

Start script is a script which gets executed before the graph execution starts.

If we want to export values of parameters to the graph then we can write in start script then run the graph then those values will be exported to graph.

How will you test a dbc file from command prompt?

Answers:

A .dbc file can be tested using **m_db** command

eg: m_db test .dbc_filename

Can we merge two graphs?

Answers:

You can not merge two ab-Initio graphs. You can use the output of one graph as input for another. You can also copy/paste the contents between graphs.

Explain the differences between api and utility mode?

Answers:

api and Utility are Database Interfaces.

api use SQL where table constraints are checked against the data before loading data into Database.

Utility uses Bulk Loading where table constraints are disabled first and data loaded into Database and then table constraints are checked against data.

Data loading using Utility is faster when compared to Api. if a crash occurs while loading data into database we can have commit and rollback in Api but we need to load whole in Utility mode.

How to Schedule Graphs in Ab-Initio,like work flow Schedule in Informatica? And where we must use Unix shell scripting in Ab-Initio?

Answers:

We can use **Autosys**, **Control-M**, or any other external scheduler to schedule graphs in Ab-Initio.

We can take care of dependencies in many ways. For example, if scripts should run sequentially, we can arrange for this in Autosys, or we can create a wrapper script and put there several sequential commands (nohup command1.ksh & ; nohup command2.ksh &; etc). We can even create a special graph in Ab-Initio to execute individual scripts as needed.

What is Environment project in Ab-Initio?

Answers:

Environment project is a special public project that exists in every Ab-Initio environment. It contains all the environment parameters required by the private or public projects which constitute AI Standard Environment.

What is Component Folding?What is the use of it?

Answers:

Component Folding is a new feature by which Co>operating System combines a group of components and runs them as a single process.

Component Folding improves the performance of graph.

Pre-Requirements for Component Folding

- The components must be foldable.
- They must be in same phase and layout.
- Components must be connected via straight flows.

Components that Foldable are:

- Reformat Re DEFINE Format
- Join (in-memory only) Replicate

Ab-Initio Interview Questions For Beginners

- | | |
|---------------------------|--------------------------|
| • Filter by Expression | Rollup |
| • Scan | Sort |
| • dedup Sorted | Sort within Groups |
| • Join With DB | Trash |
| • Normalize | Denormalize sorted |
| • Partition by Expression | Partition by key |
| • Broadcast | Partition by Round Robin |
| • Interleave | Gather |
| • Generate Records | |

How do you Debug a graph ,If an error occurs while running?

Answers:

There are many ways to debug a graph. we can use

- [Debugger](#)
- [File Watcher](#)
- [Intermediate File](#) for debugging purpose.

What do u mean by \$RUN?

Answers:

This is parameter variable and it contains only path of project sandbox run directory. Instead of using hard-code value to use this parameter and this is default sandbox run directory parameter.

fin -----> top-level directory (\$AI_PROJECT)

| |---- mp -----> second level directory (\$MP)

| |---- xfr -----> second level directory (\$XFR)

| |---- run -----> second level directory (\$RUN)

| |---- dml -----> second level directory (\$DML)

What is the importance of EME in ab-Initio?

Answers:

EME is a repository in Ab-Initio and it used for check-in and checkout for graphs also maintains graph version.

EME is source code control system in Ab-Initio world. It is repository where all the sandboxes

related(project related codes(graphs version are maintained) code version are maintained , we just check-in and checkout graphs and modified it according. There will be lock put once it is access by any users.

What is the difference between sandbox and EME, can we perform check-in and checkout through sandbox/ Can anybody explain check-in and checkout?

Answers:

Sandboxes are work areas used to develop test or run code associated with a given project.

Only one version of the code can be held within the sandbox at any time. The EME Data-store contains all versions of the code that have been checked into it.

A particular sandbox is associated with only one Project where as a Project can be checked out to a number of sandboxes.

What is difference between sandbox parameters and graph parameters?

Answers:

Sandbox Parameters are common parameters for the project. it can be used to accessible with in a project. The graph parameters are uses with in graph but you can't access outside of other graphs. It's called local parameters.

How do you connect EME to Ab-Initio Server?

Answers:

There are several ways of connecting to EME

- Set AB_AIR_ROOT
- GDE you can connect to EME data-store
- login to eme web interface
- using the air command, i don't know much about this.

What is use of co>operating system between GDE and Host?

Answers:

The co>operating system is heart of GDE, It always referring the host setting, environmental variable and functions while running the graphs through GDE. It's interfacing the connection setting information between HOST and GDE.

What is the use of Sandbox ? What is it.?

Answers:

Sandbox is a directory structure of which each directory level is assigned a variable name, is used to manage check-in and checkout of repository based objects such as mp, run, dml, db, xfr and sql (graphs, graph ksh files, wrapper scripts, dml files, xfr files, dbc files, sql files.)

Fin -----> top-level directory (\$AI_PROJECT)

| |---- mp -----> second level directory (\$AI_MP)

| |---- xfr -----> second level directory (\$AI_XFR)

| |---- run -----> second level directory (\$AI_RUN)

| |---- dml -----> second level directory (\$AI_DML)

Sandbox contains various directories, which is used for specific purpose only. The mp directory is used for storing data mapping details about between sources and targets or components and the file extension must be *.mp. The xfr directory denotes purpose of stores the transform files and the file extension must be *.xfr. The dml directory is used for storing all meta-data information of data with Ab-Initio supported data types and the file extensions must be *.dml.

The run directory contains only the graph's shell script (korn shell script) files that are created after deploying the graph.

The sandbox contains might be stores all kinds of information for data.

What is mean by EME Data Store and what is use of EME Data Store in Enterprise world?

Answers:

EME Data Store is a Enterprise Meta Environment Data store (Enterprise Repository) and its contains 'n' number of projects (sandbox) which are interfacing the meta data between them. These sandbox project objects (mp, run, db, xfr, dml) are can be easily to manage the check-in, checked out of the repository objects.

Mode:

In the EME Data-store Mode box of the EME Data-store Settings dialog, choose one of the following:

Source Code Control — This is the recommended setting. When you set a data-store to this mode, you must check out a project in order to work on it. This prevents multiple users from making conflicting changes to a project.

Full Access — This setting is strongly not recommended. It is for advanced users only. It allows you to edit a project in the data-store without checking it out.

Save Script When Graph Saved to Sandbox

In the EME Data-store Settings dialog, select this option to have the GDE save the script it generates for a graph when you save the graph. The script lets you run the graph without the GDE if, for example, you relocate the project.